How to Search the Internet Archive Without Indexing It

Kanhabua, Nattiya; Kemkes, Philipp; Nejdl, Wolfgang; Nguyen, Tu Ngoc; Reis, Felipe; Tran, Nam Khanh

Published in:
Research and Advanced Technology for Digital Libraries

DOI (link to publication from Publisher):
10.1007/978-3-319-43997-6_12

Publication date:
2016

Document Version
Early version, also known as pre-print

How to Search the Internet Archive Without Indexing It

Nattiya Kanhabua[1], Philipp Kemkes[2], Wolfgang Nejdl[2]
Tu Ngoc Nguyen[2], Felipe Reis[2], Nam Khanh Tran[2]

[1]Department of Computer Science, Aalborg University, Denmark
[2]L3S Research Center / Leibniz Universität Hannover, Germany
[1]nattiya@cs.aau.dk [2]{kemkes, nejdl, tunguyen, reis, ntran}@L3S.de

ABSTRACT

Significant parts of our cultural heritage are produced on the Web in recent years. While the easy accessibility to the current Web is a good baseline, optimal access to the past of the Web faces several challenges. This includes dealing with large-scale web archive collections, as well as lacking of usage logs, which contain implicit human feedback most relevant for today's web search. In this paper, we propose an entity-oriented search system to support retrieval and analysis processes on web archives. We use Bing, searching the current Web, to retrieve a ranked list of results, and we link our search results to the WayBack Machine; thus allowing keyword search on the Internet Archive without processing and indexing its raw content. Our system complements existing web archive search tools through a user interface, which comes close to the functionalities of modern web search engines (e.g., keyword search, query auto-completion and related query suggestion), plus the huge benefit of taking user feedback on the current Web into account also for Web Archive search. Through extensive experiments, we conduct quantitative and qualitative analyses in order to provide insights that enable further research on and practical applications of web archives.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval models*

General Terms

Algorithms, Experimentation, Performance

Keywords

Web Archive Search, Entity-based Queries

1. INTRODUCTION

Traditional institutions keeping our cultural heritage need to be complemented with facilities for preservation and public access of online cultural assets. This is critical given that even for the presumably interesting resources shared through social media like Twitter were estimated that 27% of those are lost and not archived after $2\frac{1}{2}$ years [15]. National and international initiatives have recognized this need and started to collect and preserve parts of the Web. The Internet Archive has by far the largest web archive collection among the institutions active in web preservation, where it has collected more than 2.5 Petabyte of web content since 1996. Another important European initiative is the Internet Memory Foundation, active in several EU-funded research projects on web archiving, with a set of smaller crawls for specific topics, domains and projects. Two important national libraries engaged in web preservation are the British Library and the German National Library, with the aim to preserve national web content.

Easy access to historical web information becomes more and more important as the means for accessing and exploring these archives, but the current facilities are severely underdeveloped [3, 6]. None of the archive initiatives is able to provide their collections through an interface, which comes close to the functionalities we see on today's web search engines. The Wayback Machine [18] provides the ability to retrieve and access web pages stored in the Internet Archive, requiring users to represent their information needs by specifying the URLs of Web pages to be retrieved. Any more sophisticated search and exploration is not supported and takes a lot of manual effort, contrasting sharply with the easy accessibility to the current Web through Google, Yahoo!, Bing and other Web search engines. Clearly, more sophisticated retrieval models and interfaces are needed to exploit the information, which is stored in the Internet Archive and through related efforts in other countries.

Figure 1 illustrates the top-ranked results from Bing for the query Angela Merkel, issued on August 20, 2015. The search results for a popular entity like the German politician consist of both long-term relevant URLs, e.g., a Wikipedia page or bibliography pages in news websites, as well as short-term relevant web pages, e.g., news articles. No news articles aged over one day appear in the top results, the rest of the results are long-term relevant pages, with a URL static for a long time.

A major problem with web archive search is an absence of query logs. Without search logs in web archives, it is difficult to provide a good ranking of search results without bias. To compensate this shortcoming, we built a prototype archive search system on top of Bing, which already provides a good mix of long-term and short-term relevant results. On the current web search, there are certain types of query intents that are similar to information needs on web archive search. In particular, we are interested in supporting web archive searches for *named entity queries* (or entity queries), which represent a significant fraction of current web search queries [13, 20]. For instance, when a user searches for 'Alan Turing', even though the intent is triggered by a recent movie 'The Imitation Game', the user will want to know *everything* about Alan Turing when searching a Web archive. These are the results we are aiming for, optimizing precision of our prototype, instead of recall.

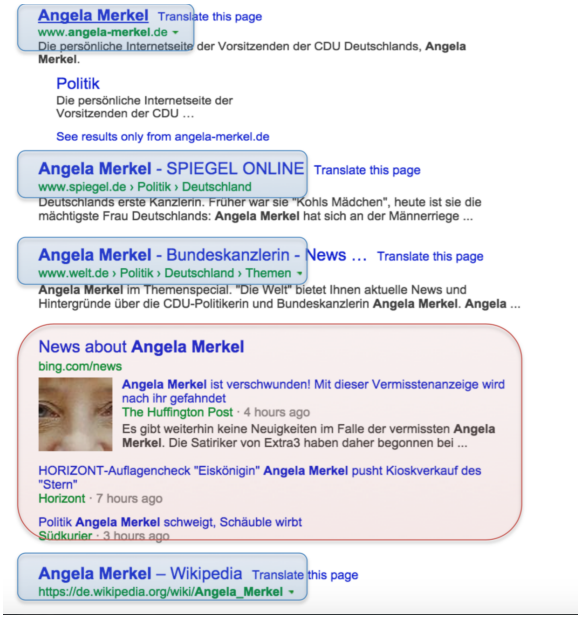

Figure 1: Top-ranked results from Bing for the query Angela Merkel. Blue areas indicate URLs that are *long-term* relevant, while the red area depicts *short-term* relevant URLs.

The goal of this work is to provide a scalable and responsive prototype to searching entity-related information on web archive collections. We propose a novel web archive search system by leveraging current web search engines and the Internet Archive. Relying on the current web search engine technologies for accessing web archives help us to achieve high quality ranking results based on search sessions and implicit human feedback. While providing entity-based indexing of Web archives is crucial, we do not address the indexing issue in this work, but instead extend the WayBack Machine API in order to retrieve archived content. This for the first time provides entity-oriented, multilingual searches on the Internet Archive, as the basis for a new kind of access to web archives.

Our contributions can be summarized as follows.

- We propose a novel system in order to support entity-based, multilingual searches as the basis for a new kind of access to Web archives.
- We make our search system publicly available for the community that enable further research on and practical application for web archives.
- Through extensive experiments, we conduct qualitative and quantitative studies and provide detailed analysis on the results returned through our Web archive search prototype.
- Finally, we outline the next steps towards more advanced retrieval and exploration of web archive content.

The rest of the paper is organized as follows. We present the problem statement in Section 2 and our approach in Section 3. Section 4 describes our analysis results, followed by a discussion in Section 5. We describe related work in Section 6 and conclude our paper in Section 7.

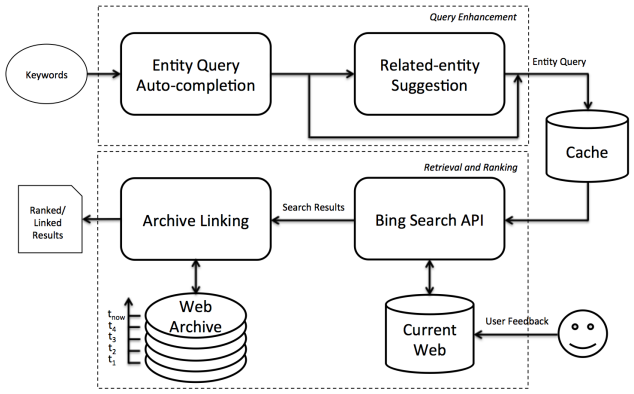

Figure 2: Our proposed system framework

2. PROBLEM STATEMENT

We describe our web archive search problem as follows.

Information needs. Recent studies have shown that *named entity queries* (or simply *entity queries*) comprise a significant fraction in search logs of the current Web [13, 20]. Due to a lack of web archive search logs, we assume similar behavior for web archive searchers, i.e., their information needs are either (1) exploring entity-related information, or (2) seeking for information related to a specific event in which entities involve. In our work, we allow users to search using entity queries that exist in Wikipedia articles in both English and German. For a given query, we determine an entity category using a heuristic approach as presented in [9].

Search results. In the context of web archives, we define two main types of web pages for entity queries: (1) general or static pages about the entity that do not change much over time (which are relevant over a long period of time), so-called *long-term relevant* and (2) dynamic pages such as news articles or blogs (which are only relevant for a short-time period), denoted *short-term relevant*.

Ranking. In this work, we aim at providing high precision entity search results, rather than optimizing recall. In order to achieve our goal, we build on the ranking of search results provided by Bing, given that the current Web search engines already try to provide a suitable mix of long-term and short-term relevant pages, while taking a lot of user feedback into account. Although we assume Bing returns relevant results, we still need to investigate search results for different kinds of entities in a principled way.

Based on our problem definition above, we will address two main aspects, namely, archive coverage and result overlap. The coverage of search results returned by Bing is concerned about how many of these results are archived on the Internet Archive. We hypothesize that many of these general pages about the entity are archived already in the Internet Archive, while news or recent pages will not be indexed yet. On the other hand, important news sites or domains themselves should all be archived. Another aspect to be addressed is the *temporal dynamics* of search results. We assume that the top-ranked results returned by Bing at different time points and rate of their change vary by entity categories. We will therefore also analyze result variations over time in order to gain more insights.

3. OUR APPROACH

The huge size of Web archives has not only created huge challenges for indexing them, but also increased the difficulty for users to manage their information needs. In addition, it is more difficult for users to compose a succinct and precise query, because of their temporal dimension.

Our archive search system provides keyword-based search functionalities, similar to existing web search engines. Figure 3 shows our search interface and retrieved results for the query "Angela Merkel". Users can issue an entity-based query for any entity described in Wikipedia in English and German. The system returns a ranked list of search results, which provides links to both the current page on the Web, as well as the archived versions in the Internet Archive, using blue and green links. Blue link refers to the current page on the Web. Green link refers to the archived versions in the Internet Archive. The system is publicly accessible via `http://alexandria-project.eu/archivesearch/`.

Figure 2 shows the overview of our search engine framework and system components, which comprise query auto-completion, Bing search API, archive linking, caching, and related entity suggestion. In the following, we will describe the proposed approach underlying our search system.

3.1 Query Auto-Completion

We support the query formulation process using query auto-completion. When the user types a query, we suggest a short list of relevant entities in order to help the user complete his information need. Figure 4 shows the auto-completion for the query "Angela Merkel". We use a Wikipedia entity index comprised of all Wikipedia entities and store it in a trie data structure to allow fast prefix lookup. Additionally, we split all entities at white and special characters. All strings starting at each token are added to the trie as an addtional reference to the original entity. Furthermore, we also take into account simplified versions of all tokens which contained letters with accents in our index. This allows our application to suggest entities even if the user does not know the exact name or cannot type the name in a foreign alphabet. As an example, for the query "schroder" we would suggest the former German chancellor "Gerhard Schröder". We further rank the suggested query completions by their popularity using the cumulative page views (see the detailed description below). To penalize the time-sensitive popularity of the entities, the (daily) page views are accumulated over a long period. Finally, the entity selected by the user is sent as input to the search API.

The Wikipedia page views [1] refer to the number of times a particular Wikipedia page has been requested; thus being viewed by users. In Wikipedia, the view counts for pages that redirect to a given page are not combined with page views of the page being redirected to. In this work, we aggregate all these related views to present the popularity (reflected by the page views) of an entity query for all its query variants. We computed the aggregated statistics of page views over a period of 4 years, from January 2011 to August 2015.

3.2 Bing Search API

Bing is Microsoft's search engine providing access to their current web index through a RESTful API available at the Azure Marketplace. Bing returns results in XML or JSON data formats and offers two different API endpoints, the full featured *Bing Search API*, and the restricted and less expensive *Bing Search API - Web Results Only*. The later lacks of few meta data like the overall result count. Nevertheless, it provides all basic search result information, such as URL, title and a text snippet. By specifying parameters, we can request optimized results for different languages/countries. We therefore use the endpoint with web results only. Yahoo! and Google provide similar search APIs but at higher costs and more restrictions.

[1] `https://en.wikipedia.org/wiki/Wikipedia:Pageview_statistics`

Entity	Views
Deutschland	5,702,314
Nekrolog 2014	5,530,098
Game of Thrones	5,215,327
Wikipedia	5,187,845
Chrome	5,079,345
Chris Kyle	4,622,544
The Big Bang Theory	3,092,815
Facebook	2,960,010
Charlie Hebdo	2,827,769

Table 1: Top-10 most viewed Wikipedia articles in German

3.3 Archive Linking

After obtaining search results from the Bing search API, we link the ranked list of results to the WayBack Machine to support browsing through the archived versions of web pages. The WayBack Machine is a tool provided by the Internet Archive that allows access to its web archives by specifying a URL. The URL-based access can be programmatically used through an API provided by the Internet Archive. For a given URL it returns a list of all dates when the URL has been archived. When a URL has been archived many times in the past, retrieval can take very long time. We therefore use two requests to retrieve only the first and last capture dates to display the timespan at which the URL has been archived. When the temporal intent of the user is provided, we narrow down to return only the revisions around the interested time point.

3.4 Result Caching

To avoid recurring requests to the Bing and WayBack Machine APIs, we store the search results locally in our cache, using a simple relational database. In order to take into account the fact that search results change over time, we update search results monthly to keep our cache up-to-date and to track changes at both sources. Besides the queries entered by our users, we use also the 10,000 most viewed English Wikipedia entities as queries. As a side effect, this procedure results in building a corpus of past search results, which will support promising, longitudinal studies of web archive search, investigating how results change over time, triggered by events or changing user behavior. In Section 4, we will analyze the cache in order to reveal several important aspects, e.g., how long a web page stays relevant and when it fades away from top-ranked results. These insights will help improving the next version of our Web archive search prototype.

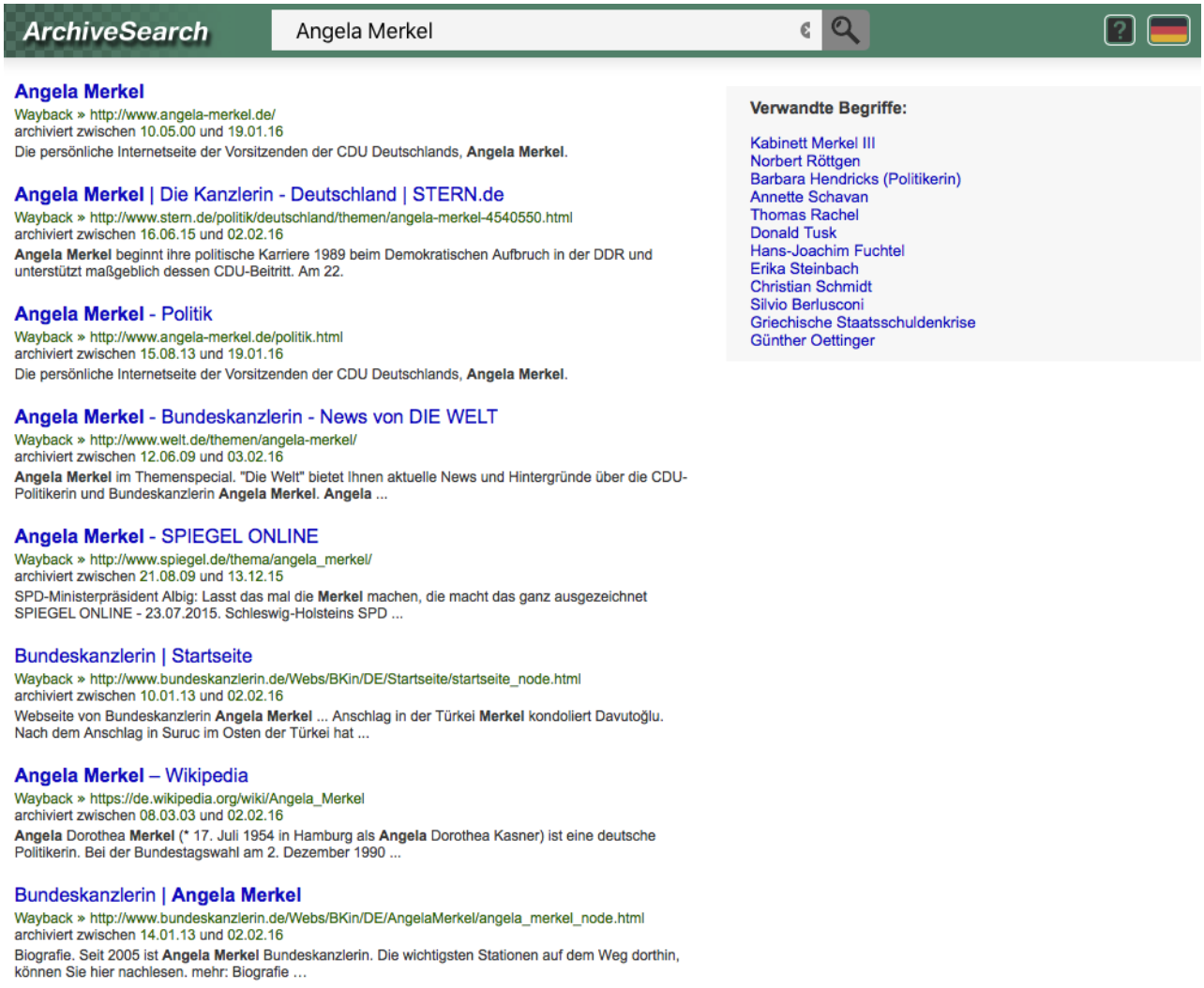

Figure 3: Search results for Angela Merkel search in German

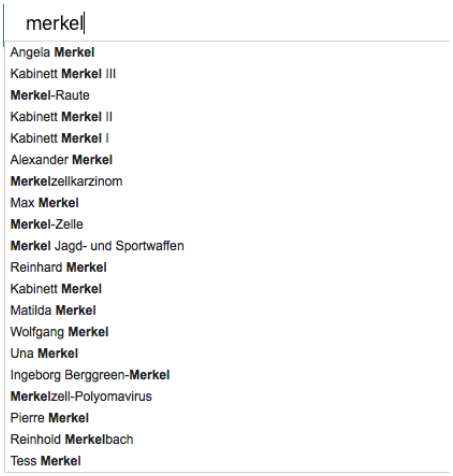

Figure 4: Auto-completion for the query "Angela Merkel" in German search

3.5 Related Entity Suggestion

A traditional web search engine supports exploration by suggesting related queries, which is based on analyzing search sessions and identifying co-occurrences of the issued query. For web archive search, we do not have query logs for obtaining search sessions, thus we leverage a dump of Wikipedia articles and build an entity graph in order to find related queries for our entity-oriented search. We follow the approach to determining the link-based entity relatedness originally proposed in [14].

Relatedness between two entities e_1 and e_2 is measured based on the overlap between the set of Wikipedia articles that link to e_1 and the set of articles that link to e_2.

$$relatedness(e_1, e_2) = \frac{log(max(|S_1, S_2|)) - log(|S_1 \cup S_2|)}{log(|W| - log(min(|S_1, S_2)} \quad (1)$$

whereas S_i is the set of articles that links to entity e_i, and W is the set of all articles in Wikipedia.

Our entity graph is constructed from a recent Wikipedia dump (downloaded from September 2015 for both English and German Wikipedia), with the assumption that the link-based relationships between a pair of two entities are accumulated; thus relaxing time sensitivity (in this case, the relatedness does not bias to any time point). An interesting extension will be to take time dependent relationships into account, planned for the next version of our system.

3.6 Multilinguality

The described components are designed to support different languages. Dependent on the language selected by the user, we use the corresponding Wikipedia version to generate the query auto-completion and the related entities suggestions. More importantly, we request search results from Bing which are optimized for the selected language and region. Furthermore, we leverage Wikipedia inter-language links when a user changes the front-end language. For example: When an English user searched the term "climate change" and switches to German, he will be redirected to "Klimawandel". Currently, we support English and German, but will add additional languages in the future.

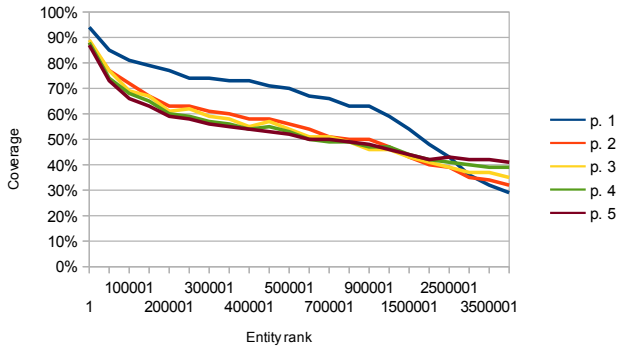

Figure 5: Coverage (percentage) of archived content at different top-k results over entities ranked by their popularity

4. EXPERIMENTS

We conducted extensive experiments in order to gain insights into our assumptions presented in Section 2. We seek to answer two main research questions as follows.

RQ1. What is the coverage of archived content provided by the current web search engine?

RQ2. To what extent the search results change over time, and why they change?

In the following, we will divide our experimental results into two main parts, where we describe our quantitative and qualitative analyses for each of the aforementioned research questions.

Part I: Analysis Results for RQ1.

Our system relies on the assumption that many pages returned as search results for entity queries are archived by the Internet Archive. To check this assumption, we took all English Wikipedia entities sorted by their view count and selected buckets of 1,000 entities at different positions in this list to represent entities from different popularity categories. We started with the 1,000 most viewed entities and continued with the entities from position 50,001 to 51,000 and so on. For each entity in an individual bucket, a search query was conducted and we checked how many results on the first five pages (10 results per page) were archived by the Internet Archive. Table 2 and Figure 5 show the average results per page and bucket.

The results of popular entities (rank: 1 - 1000) show a very high coverage with the Internet Archive. On page one, 94% of the results are archived. On pages two to five, still 87% to 89% are available at the Internet Archive. Overall the coverage declines for less popular entities. Interestingly, it drops faster for the first pages of the search results than for the posterior ones. Upon inspection, this seems to be caused by the fact that Bing ranks recent results higher (for example on the first page of its search results), while the Internet Archive needs much more time to archive less popular pages, if occurred.

To gain further insights, we conducted a coverage study by entity category. More precisely, we analyzed top-100 search results for 300 popular entities from 14 different categories,

Entity rank	p. 1	p. 2	p. 3	p. 4	p. 5
1 - 1000	94%	88%	89%	88%	87%
50001 - 51000	85%	77%	77%	74%	73%
100001 - 101000	81%	72%	69%	68%	66%
150001 - 151000	79%	67%	67%	65%	63%
200001 - 201000	77%	63%	61%	60%	59%
250001 - 251000	74%	63%	62%	59%	58%
300001 - 301000	74%	61%	59%	57%	56%
350001 - 351000	73%	60%	58%	56%	55%
400001 - 401000	73%	58%	55%	54%	54%
450001 - 451000	71%	58%	57%	55%	53%
500001 - 501000	70%	56%	54%	53%	52%
600001 - 601000	67%	54%	51%	50%	50%
700001 - 701000	66%	51%	51%	49%	50%
800001 - 801000	63%	50%	49%	49%	49%
900001 - 901000	63%	50%	46%	47%	48%
1000001 - 1001000	59%	47%	46%	47%	46%
1500001 - 1501000	54%	43%	43%	44%	44%
2000001 - 2001000	48%	40%	41%	42%	42%
2500001 - 2501000	43%	39%	39%	41%	43%
3000001 - 3001000	36%	35%	37%	40%	42%
3500001 - 3501000	32%	34%	37%	39%	42%
4000001 - 4001000	29%	32%	35%	39%	41%

Table 2: Coverage (percentage) of archived content at different top-k results over entities ranked by their popularity

for example, actor, journalist, painter, and politician, where we analyzed approximately 20 entities per category. We only considered search results with .DE domains (German web pages), and checked the coverage with our local German web archive[2], instead of web archives of the Internet Archive. As shown in Figure 6, the coverage statistics on the German web archive shows significantly lower results than the one based on comparision with the Internet Archive due to search result bias towards English web pages, in general, even for the German version of Bing. Categories with lower coverage tend to associate with recent and dynamic web content, whereas the results of the categories with higher coverage are rather static and less changed. Note that result URLs are not always archived (e.g. for newspaper articles), but nearly all domains (i.e. news sites) are archived, regardless of entity categories.

We also performed search result annotation for 9 entities, where we employed 5 human assessors to manually label top-100 search .DE results (filtered out non .DE domains). For each (query, URL) pair, we asked at least 4 assessors to give a label based on relevance assessment criteria consisting of three scales: *long-term relevant*, *short-term relevant* and unknown. The results are shown in Table 3, where we can notice that non-active entities, such as, Pablo Picasso and Ernest Hemingway, have more long-term relevant results than active entities like Elon Musk and Leonard Nimoy (to be expected).

In the following, we provide detailed analyses for selected entities in order to better understand the coverage aspect of search results and web archives.

[2]It contains the entire German domain (.de) from 1996 to 2013, which is provided by the Internet Archive. The size of this collection is about 36TB of compressed textual data and 80TB when including images and multimedia data. In total, the German collection consists of 5.51 billion captures (with duplicates) and contains 2.3 TB uncompressed metadata (CDX).

Entity	Category	Long-term	Short-term
Leonard_Nimoy	Actor	52%	48%
Elon_Musk	Business people	37.50%	62.50%
Costa_Concordia_disaster	Incidents	52.40%	47.60%
Ernest_Hemingway	Journalist	81.48%	18.52%
Giuliana_Rancic	Journalist	40%	60%
Pablo_Picasso	Painter	97.60%	2.40%
Banksy	Painter	48.10%	51.2%
Vietnam	Politics	100%	0%
Ku_Klux_Klan	Politics	12.50%	87.50%

Table 3: Percentage of long-term relevant and short-term relevant pages in top-100 results (filtered out non .DE domains)

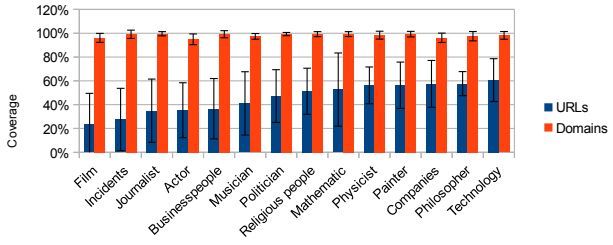

Figure 6: Coverage (percentage) of archived URLs and domains at top-100 results for different entity categories

Lady Gaga (view count rank: 108; views: 9.453.966; querying date: 19.01.2016).

Almost everybody on the Web knows Lady Gaga. She is an American songwriter, singer, actress, philanthropist, dancer and fashion designer, famous and higly popular since 2009. Her Bing top-50 results have almost complete coverage on the Internet Archive. More specifically, 98% of the results are archived. Among the results, only one URL[3] is not archived. This URL points to a gossip news inside the entertainment section of the New York Daily News web site. The relevance of this article is rather low as the main content of this new is about Linda Perry, another entity from the music domain, who made some aggressive comments about Lady Gaga during an Oscar nomination. We checked again on January 25, 2016, and the URL was still not archived. The URL was published online on January 18, 2016.

Pablo Picasso (view count rank: 479; views: 5.386.218; querying date: 19.01.2016).

Pablo Ruiz y Picasso (better known as Pablo Picasso) was a Spanish painter, sculptor and poet. He is known to be one of the greatest and most influential artists of the 20th century. For him, 94% of the top 50 Bing results are archived. Given his importance and popularity, high Internet Archive coverage was expected. Still, a few results are not covered by the Internet Archive. These nonarchived URLs can be classified as: category

[3]`http://www.nydailynews.com/entertainment/gossip/linda-perry-slams-lady-gaga-article-1.2500319`

search for products on sales on ebay.com[4] - this URL show s products related to the entity like paintings, drawings, and books; news website topic search[5] - showing news related to the entity published on the New York Times. As Pablo Picasso has been dead for 48 years, he is not mentioned a lot on social networks and news. Once more, we see URLs related to news or products not being archived, but we do have archived URLs under the products category, for instance an amazon.com URL[6] showing products related to Picasso. No news websites are included for the top 50 Bing results, which is plausible. On the other hand, URLs pointing to static pages - like Wikipedia and biography specialized websites like pablopicasso.org - are very likely to be archived and have the highest ranking. Bing returns three language-based URLs (pt - Portuguese, fr - French, and de - German) all of them point to the same Wikipedia content, archived as well, but somewhat redundant.

Direct metal laser sintering (view count rank: 150.849; views: 108.275; querying date: 18.01.2016).

Direct metal laser sintering (DMLS) is an additive manufacturing technique that uses a Yb (Ytterbium) fibre laser fired into a bed of powdered metal, aiming the laser automatically at points in space defined by a 3D model, melting or rather, welding the material together to create a solid structure. DMLS was developed by the EOS company based in Munich, Germany. This is an interesting example for a complex technical process. which explains the low rank and views count rates. The percentage of archived results is 84%, though, still very high. Among the non-archived URLs are: a white paper[7] about the materials used for DMLS in .PDF format; an website[8] from a company which is supplier of mechanical parts - presenting the materials used for DMLS; a non-available page[9] from a 3D printing company website; and a page[10] from a 3D printing and manufacturing company - giving the reader an overview on what is the technique about. These pages are all rather static and were published months or years ago but, in comparison with other archived results for this and others entities, they have too low web traffic in order to be archived. Among the archived results, we have other 3D printing companies websites like a company website from Florida, US; a Wikipedia page; and also a youtube.com video showing the process from the 3D model creation till the printed part. For this last example, the youtube.com views count were 317,773 on February 10, 2016. No news are included in the top 50 results.

Battle of Rathmines (view count rank: 1.000.798; views: 8.723; querying date: 18.01.2016).

The Battle of Rathmines was fought in and around what is now the Dublin suburb of Rathmines in August 1649, during the Irish Confederate Wars, the Irish theatre of the Wars of the Three Kingdoms. The number of Wikipedia page views for this entity which is 8723, not very many people know about it. Still, about 56% from the top 50 Bing results are archived. Among the other 44% which is not archived there is a page[11] from a website that shows data from Wikipedia and it is number 4 by Bing, and another one, at postion 48, which is a post[12] from a blog. This entity is related to a real event in Ireland history that took place in almost 500 years ago, but is very local and rather unimportant outside of Ireland.

Ellison Quirk (view count rank: 4.000.425; views: 464; querying date: 18.01.2016).

Ellison Quirk was born in 1866 in the town of Lucknow, near Orange, New South Wales on the Wentworth goldfields. He eventually established himself within the Manly community as a Storekeeper and Land Agent. Quirk stood for election to the first council of the newly proclaimed Warringah Shire Council on 3 December 1906. Subsequently elected as an Alderman, he rose to be Shire President on three consecutive occasions in 1910, 1913-1914 and 1918-1919. Its 464 Wikipedia page views testify to the fact, that almost nobody is interested in Ellison Quirk. For him, Bing also returns results which are not relevant, such as the one for Robert Quirk[13]. Archiving rate for Quirk is only 32%, and among the archived results we have non-related results, such as a Twitter account URL[14] ranked in position 4 by Bing, and with only about 20% of the top-50 results really related to Ellison Quirk.

Part II: Analysis Results for RQ2.

In this section, we present the analysis of Bing results for entity-based queries executed in three time periods: June 2015, August 2015 and January 2016. For a given entity, we compute the overlap between the top-ranked results at different time periods. Through this, we gain insights into our question RQ2, on how Bing query results change over time. We conducted a study on 300 popular queries of 14 different categories as explained in RQ1. We discuss a few samples ranging from low to high overlapping rates.

Figure 7 illustrates the change of search results (as measured by the overlap statistics) over time for different entity categories over the period of 2,5 and 7 months, respectively. In general, result change after 2 months results in approximately 47% of the top-100 URLs not returned any more. After 7 months, result change increases to 60%. Across different categories, there is not much difference in temporal dynamics for the search results. The *Actor* category varies most, whereas the result variation is least for *Philosopher*. The main explanation here is that the entities in our *Actor*

[4] http://www.ebay.com/sch/i.html?_nkw=pablo+picasso
[5] http://www.nytimes.com/topic/person/pablo-picasso
[6] http://www.amazon.com/Pablo-Picasso/e/B001H6SD9G
[7] http://s3.amazonaws.com/cloudfab_bukkit/the_files/73/original/EOS_DMLS.pdf?1255724232
[8] https://www.anubis3d.com/technology/direct-metal-laser-sintering/materials/
[9] http://www.bastech.com/3dproducts/production-3dprinters/direct-metal-laser-sintering/
[10] https://www.stratasysdirect.com/solutions/direct-metal-laser-sintering/

[11] http://www.thefullwiki.org/Battle_of_Rathmines
[12] http://irelandinhistory.blogspot.de/2014/08/blog-post_11.html
[13] http://www.myheritage.com/names/robert_quirk
[14] https://twitter.com/ellisonbarber

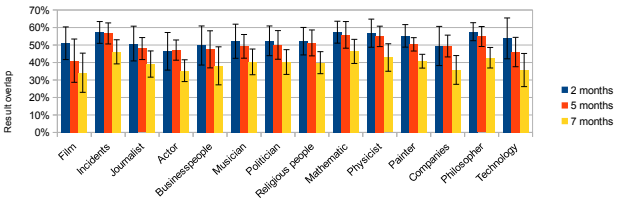

Figure 7: Result overlap for different entity categories

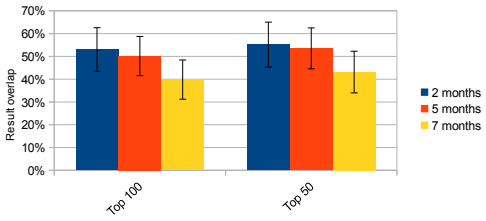

Figure 8: Result overlap for top-50 and top-100 results

category are more active than entities in the *Philosopher* category, thus having more short-term relevant queries in the top-100 result list, which change over time. We observe in Figure 8 that the overlap in the top-50 result is slightly higher than in the top-100 result. This indicates there are less result changes in the first 50 results than in the next 50 results. However, the difference is not significant.

Entity: Frida Kahlo

Frida Kahlo was a Mexican painter known for her self-portraits. Kahlo's life began and ended in Mexico City, in her home known as the Blue House. Her work has been celebrated in Mexico as emblematic of national and indigenous tradition, and by feminists for its uncompromising depiction of the female experience and form. Bing search results for this entity feature a high number of biography pages like fridakahlo.org and commercial places which took her name like the Mexican food restaurant fridakahlo.de. Results contain only a small number of news, some articles on sale, for instance paintings on ebay.com and also a few blog posts. Half of the top 50 results in August overlap the top 50 results from June. For the top hundred during the same period the percentage is quite similar being 46%. In January 2016, the overlap with the June results the overlap is 38% for the top 50 and 36% for the first 100 results. Results which do not reappear in later months include blog posts, articles for sale and non-related entities for instance the Mexican restaurant previously mentioned. Recent events for this entity are rare as Frida Kahlo is already dead so we don't have regular mentions on news websites. Nevertheless, blog posts mentioning her are ranked high. Blog posts URLs have a decreasing rank from June to August, though, and some are not shown in the January result set anymore.

Entity: Barack Obama

Barack Obama's results include URLs pointing to news, topic search themes or categories in web portals like ndtv.com, books for sale on Amazon, and biography or permanent pages like Wikipedia and official pages like barackobama.com. 47.5% of the results from August overlap with the results retrieved in June. These shared results are URLs pointing to biography or permanent pages and most of them appear with a high rank. The remaining 52.5% are URLs mostly to news or categories inside news web portals, which change over time. Overlapping of search results in tihs context decreases proportional to the number of recent events related to the entity: the higher the proportion of news results, the lower the overall overlapping with future resultsets.

Entity: Max Planck

Max Planck was a German theoretical physicist whose work on quantum theory won him the Nobel Prize in Physics in 1918, ad who did in 1947. Most results are biography pages, for example on the Nobel Prize website, and pages related to other entities which took his name, for instance the Max Planck Institute for Biology of Ageing. Total overlap is 50%, explainable by his death a long time ago and few news results.

Entity: Donald Trump

Donald Trump is currently running as a United States presidential candidate, for the election in 2016. Because of his outspoken / rude manner and his colorful career, he is well known, even though less so when we first collected results for him in 2015. For Donald Trump, more than 80% of the URLs are news (National Journal, newsobserver.com, NBC). Most of these news where published in June and don't overlap in August and January - data analyzed in this experiment shows that the URLs from the category news which is listed in results from June will probably not be shown in August or January. Total overlap average is 24%, which is very low. Even for the shorter periods, June to August and August to January, overlap is low with 30%. Search results for Trump are a very clear sample for fast changing results, caused by frequent news articles.

5. DISCUSSION

Although the system described in this paper already provides interesting functionalities, it is obviously still work in progress. As one important extension of functionality, we are working on more complex types of entity-based queries in order to support exploratory search, e.g., giving a main entity Donald Trump and related search intents. The search intent can consist of an entity, such as, *Hillary Clinton* aiming to find all events which involve these two entities, and a specific time period such as *2015-2016* narrowing down search results to a specific time period, or any contextual query, such as a concept *presidential campaign*. Figure 9 shows an example of such an exploratory query. Using this functionality, we can explore different aspects of a given entity in web archives in different and meaningful ways.

Another important aspect for our future development is to advance our ranking. As our current method is relying on the Bing search API, we sacrifice recall. Learning from Bing over time as well as from our user logs, we will be able to provide more sophisticated ranking taking different features into account. Bing results can act as 'soft' ground truth for learning the model. Our ranking model will then be able to return relevant documents which are not longer available on

the current Web.

Finally, we also work on improving our suggestion components, i.e., related entity suggestion to deal with queries having time as another aspect (e.g., Obama 2008). In the current system, we exploited a state-of-the-art method to suggest related entities to the query entity, with the assumption that their relationship strengths are accumulated over time. This relationship measure is reasonable to serve for queries with arbitrary relevant time. However, in reality relationships between entities do change over time, typically triggered by events. We can therefore return different related entities for different time periods to the input entity. For instance, the entities mostly related to *Hillary Clinton* in 2008 should differ from those in 2012, because of her different political positions. Moreover, with an exploratory query for example *Donald Trump* and a search intent *Hillary Clinton*, it is more helpful to recommend the entities which are related to both *Donald Trump* and *Hillary Clinton*.

6. RELATED WORK

In this section, we outline related work including the existing systems that support access on web archives. We also describe current state-of-the-art for search and exploration on temporal document collections.

6.1 Existing search and access support in Web archives

An enormous amount of information is stored in Web archives. To date, there are a few search prototpyes to provide acces to these archives, but all with a number of limitations. In 2009, the Internet Archive ran a pilot in providing full-text searchability for parts of their archive, making the first five years of their archive (1996-2000) available for searching. The search ranking mechanisms available at that time were not adequate, however, and the search results were full of spam. Current tools for supporting search and exploration in Web archives are limited [4] as well. Some additional projects exist that provide limited support for archival Web research.

The Wayback Machine: The Wayback Machine is a web archive access tool supported by the Internet Archive. The Internet Archive is a non-profit organization with the goal of preserving digital document collections as cultural heritage and making them freely accessible online. The Wayback Machine provides the ability to retrieve and access web pages stored in a web archive, but it requires a user to access data by specifying the URL of a web page to be retrieved. For example, given the query URL `http://www.usa.gov`, the results of retrieval are displayed in a calendar view which displays the number of times the URL was crawled by the Wayback Machine (not how many times the site was actually updated). is not possible to search by keywords.

Memento Project: The Memento project [15] is a United States National Digital Information Infrastructure and Preservation Program (NDIIPP)-funded project aimed at making Web-archived content more readily discoverable. Rather than expecting people to know about the growing number of Web archives, and to guess which archive might hold an older version of the resource they're looking for, Memento proposes to make archived content discoverable via the original URL that the searcher already knew about, and redirecting the searcher to the archive which hosts the page at the time indicated by the user.

Archive-IT: Archive-IT [16] is web archiving service for collecting and accessing cultural heritage sites on the web, built by the Internet Archive. The service supports organizations to harvest, build, and preserve collections of digital content. Its functionalities include full-text search on the archived collections.

6.2 Search result exploration and exploratory search

Recent work on visualizing search results using a timeline has been proposed to give an overview of possible time periods relevant to the query and provide these as a hint to the user [1, 2]. Matthews et al. [12] present Time Explorer, a search engine that allows users to see how topics have evolved over time and how they might continue to evolve in the future. In the context of searching Web archives, our project will support the task of exploratory search [11, 19], also known as general information seeking. The focus of exploratory search is on human-computer interaction, helping users to complete search activities by focusing on the involvement of users and the interaction between users and systems. Research suggests that an exploratory search system should support features such as: (1) query formulation [8] (2) the leveraging of contextual information [5, 10], (3) faceted search presentation [16], (4) customizable visualizations [17], and (5) learning skills and tools for understanding knowledge [7, 14]. Contextual information includes time, place, the history of interaction, or current situations about a user. Contextual information can be captured by asking users to mark relevant results, or to indicate useful text fragments.

7. CONCLUSION

In this paper, we proposed a web archive search prototype that for the first time supports entity-oriented queries on the Internet Archive. Our system leverages Bing and the WayBack Machine to allow users to search the past Web. We provided search functionalities including keyword search, query auto-completion, query suggestion, and a ranked list of results, which are close to the current search engine systems. We conducted extensive analyses that shed light on web archive search. In addition, we included a longer discussion of future work as well as insights on the ideas/challenges for the next steps for our Web archive search investigations.

References

[1] O. Alonso, K. Berberich, S. Bedathur, and G. Weikum. Neat: News exploration along time. In *Proceedings of the 32nd European conference on Advances in Information Retrieval*, pages 667–667. Springer-Verlag, 2010.

[2] O. Alonso, M. Gertz, and R. Baeza-Yates. Clustering and exploring search results using timeline constructions. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 97–106. ACM, 2009.

[3] M. Costa, D. Gomes, F. Couto, and M. Silva. A survey of web archive search architectures. In *Proceedings of the 22nd International Conference on World Wide Web (Companion)*, WWW '13, pages 1045–1050, 2013.

[15] `http://timetravel.mementoweb.org`

[16] `https://archive-it.org`

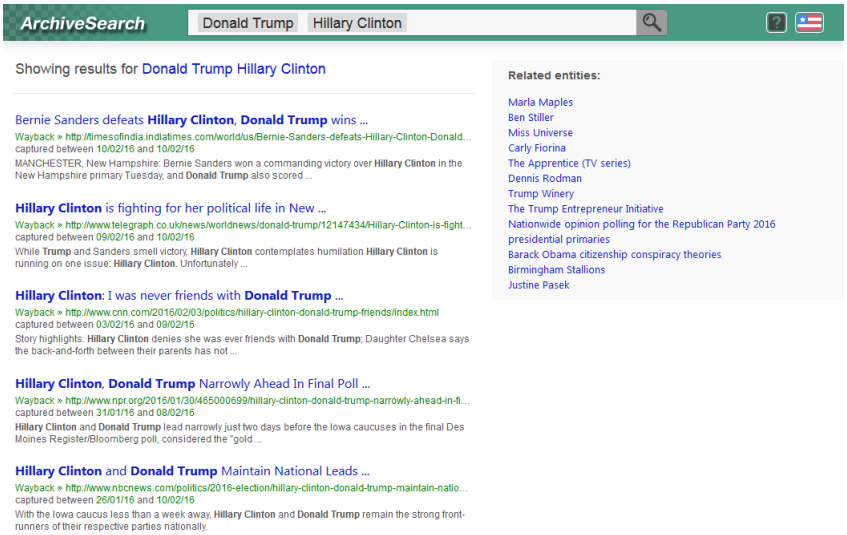

Figure 9: Example of exploratory query within our system

[4] M. Dougherty and C. van den Heuvel. Historical infrastructures for web archiving: Annotation of ephemeral collections for researchers and cultural heritage institutions. 2009.

[5] S. Dumais, E. Cutrell, R. Sarin, and E. Horvitz. Implicit queries (iq) for contextualized search. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 594–594. ACM, 2004.

[6] D. Gomes, J. a. Miranda, and M. Costa. A survey on web archiving initiatives. In *Proceedings of the 15th International Conference on Theory and Practice of Digital Libraries: Research and Advanced Technology for Digital Libraries*, TPDL'11, pages 408–420, 2011.

[7] J. He, M. de Rijke, M. Sevenster, R. van Ommering, and Y. Qian. Generating links to background knowledge: a case study using narrative radiology reports. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1867–1876. ACM, 2011.

[8] N. Kanhabua and K. Nørvåg. Determining time of queries for re-ranking search results. In *Research and Advanced Technology for Digital Libraries*, pages 261–272. Springer, 2010.

[9] N. Kanhabua and K. Nørvåg. Exploiting time-based synonyms in searching document archives. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, JCDL '10, pages 79–88, New York, NY, USA, 2010. ACM.

[10] D. Kelly and N. J. Belkin. Display time as implicit feedback: understanding task effects. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 377–384. ACM, 2004.

[11] G. Marchionini. Exploratory search: from finding to understanding. *Communications of the ACM*, 49(4):41–46, 2006.

[12] M. Matthews, P. Tolchinsky, R. Blanco, J. Atserias, P. Mika, and H. Zaragoza. Searching through time in the new york times. Citeseer, 2010.

[13] I. Miliaraki, R. Blanco, and M. Lalmas. From "selena gomez" to "marlon brando": Understanding explorative entity search. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 765–775, New York, NY, USA, 2015. ACM.

[14] D. Milne and I. H. Witten. Learning to link with Wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 509–518. ACM, 2008.

[15] H. M. SalahEldeen and M. L. Nelson. Losing my revolution: how many resources shared on social media have been lost? In *Proc. 2nd Int. Conf. on Theory and Practice of Digital Libraries*, pages 125–137, 2012.

[16] D. Tunkelang. Faceted search. *Synthesis lectures on information concepts, retrieval, and services*, 1(1):1–80, 2009.

[17] F. B. Viegas, M. Wattenberg, F. Van Ham, J. Kriss, and M. McKeon. Manyeyes: a site for visualization at internet scale. *Visualization and Computer Graphics, IEEE Transactions on*, 13(6):1121–1128, 2007.

[18] The Wayback Machine. `https://archive.org/web`. (Accessed 12 February 2016).

[19] R. W. White and R. A. Roth. Exploratory search: Beyond the query-response paradigm. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 1(1):1–98, 2009.

[20] X. Yin and S. Shah. Building taxonomy of web search intents for name entity queries. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 1001–1010, New York, NY, USA, 2010. ACM.